SYSTEMS AND METHODS FOR SPEAKER CHANGE DETECTION

GOVERNMENT CONTRACT

[0001] The U.S. Government may have a paid-up license in this invention and the right in limited circumstances to require the patent owner to license others on reasonable terms as provided for by the terms of Contract No. DARPA F30602-97-C-0253.

RELATED APPLICATIONS

[0002] This application claims priority under 35 U.S.C. § 119 based on U.S. Provisional Application No. 60/419,214 filed October 17, 2002, the disclosure of which is incorporated herein by reference.

BACKGROUND OF THE INVENTION

A. Field of the Invention

[0003] The present invention relates generally to speech processing and, more particularly, to speaker change detection in an audio signal.

B. Description of Related Art

[0004] Speech has not traditionally been valued as an archival information source. As effective as the spoken word is for communicating, archiving spoken segments in a useful and easily retrievable manner has long been a difficult proposition. Although the act of recording audio is not difficult, automatically transcribing and indexing speech in an intelligent and useful manner can be difficult.

[0005] Speech is typically received by a speech recognition system as a continuous stream of words without breaks. In order to effectively use the speech in information management systems (e.g., information retrieval, natural language processing and real-time alerting systems), the speech recognition system initially processes the speech to generate a formatted version of the speech. The speech may be transcribed and linguistic information, such as sentence structures, may be associated with the transcription. Additionally, information relating to the speakers may be associated with the transcription. Many speech recognition applications assume that all of the input speech originated from a single speaker. However, speech signals, such as news broadcasts, may include speech from a number of different speakers. It is often desirable to identify different speakers in the transcription of the broadcast. Before identifying an individual speaker, however, it is first necessary to determine when the different speakers begin and end speaking (called speaker change detection).

[0006] Fig. 1 is a diagram illustrating speaker change detection using a conventional speaker change detection technique. An input audio stream is divided into a continuous stream of audio intervals 101. The intervals each cover a fixed time duration such as 10ms. Thus, in the example shown in Fig. 1, in which five intervals 101 are shown, a total of 50 ms of audio is illustrated. Boundaries between intervals 101, such as boundaries 110-113, are considered to be potential candidates for a speaker change determination. The conventional speaker change detection system would examine all, or a predetermined number of intervals 101, to the left of a particular boundary, such as boundary 112, and to the right of boundary 112. If the audio to the left and right of boundary 112 was determined to be dissimilar enough, boundary 112 was assumed to be a speaker change. Whether the audio to the left and right of boundary 112 is similar or dissimilar may be based on comparing cepstral vectors for the audio.

[0007] One drawback to the conventional speaker detection technique described above is that the interval boundaries occur at arbitrary locations. Accordingly, to ensure that a speaker change boundary is not missed, the intervals are assigned relatively short interval durations (e.g., 10ms). Calculating whether audio segments are similar or dissimilar at each of the interval boundaries can, thus, be computationally burdensome.

[0008] There is, therefore, a need in the art for improved speaker change detection techniques.

SUMMARY OF THE INVENTION

[0009] Systems and methods consistent with the principles of this invention provide for fast speaker boundary change detection.

[0010] A method consistent with aspects of the invention detects speaker changes in an input audio stream. The method includes segmenting the input audio stream into predetermined length intervals and decoding the intervals to produce a set of phones corresponding to each of the intervals. The method further includes generating a similarity measurement based on a first portion of the audio stream within an interval prior to a boundary between adjacent phones and a second portion of the audio stream within the interval and after the boundary. Further, the method includes detecting speaker changes based on the similarity measurement.

[0011] A device consistent with another aspect of the invention detects speaker changes in an audio signal. The device comprises a processor and a memory containing instructions. The instructions, when executed by the processor, cause the processor to: segment the audio signal into predetermined length intervals and decode the intervals to produce a set of phones corresponding to the intervals. The instructions additionally cause the processor to generate a similarity measurement based on a first portion of the audio signal prior to a boundary between phones in one of the sets of phones and a second portion of the audio signal after the boundary, and detect speaker changes based on the similarity measurement.

[0012] Yet another aspect of the invention is directed to a device for detecting speaker changes in an audio signal. The device comprises a segmentation component that segments the audio signal into predetermined length intervals and a phone classification decode component that decodes the intervals to produce a set of phone classes corresponding to each of the intervals. Further, a speaker change detection component detects locations of speaker changes in the audio signal based on a similarity value calculated over a first portion of the audio signal prior to a boundary between phone classes in one of the sets of phone classes and a second portion of the audio signal after the boundary in the one set of phone classes.

[0013] Another aspect of the invention is directed to a system comprising an indexer, a memory system, and a server. The indexer receives input audio data and generates a rich transcription from the audio data. The rich transcription includes metadata that defines speaker changes in the audio data. The indexer further includes a segmentation component configured to divide the audio data into overlapping segments and a speaker change detection component. The speaker change detection component detects locations of speaker changes in the audio data based on a similarity value calculated at locations in the segments that correspond to phone class boundaries. The memory system stores the rich transcription and the server receives requests for documents and responds to the requests by transmitting ones of the rich transcriptions that match the requests.

BRIEF DESCRIPTION OF THE DRAWINGS

[0014] The accompanying drawings, which are incorporated in and constitute a part of this specification, illustrate the invention and, together with the description, explain the invention. In the drawings,

[0015] Fig. 1 is a diagram illustrating speaker change detection using a conventional speaker change detection technique;

[0016] Fig. 2 is a diagram of a system in which systems and methods consistent with the present invention may be implemented;

[0017] Fig. 3 is a diagram illustrating an exemplary computing device;

[0018] Fig. 4 is diagram illustrating functional components of the speaker segmentation logic shown in Fig. 2;

[0019] Fig. 5 is a diagram illustrating an audio stream broken into intervals consistent with an aspect of the invention;

[0020] Fig. 6 is a diagram illustrating exemplary phone classes decoded for an audio interval; and

[0021] Fig. 7 is a flow chart illustrating the operation of the speaker change detection component shown in Fig. 4.

DETAILED DESCRIPTION

[0022] The following detailed description of the invention refers to the accompanying drawings. The same reference numbers in different drawings may identify the same or similar elements. Also, the following detailed description does not limit the invention. Instead, the scope of the invention is defined by the appended claims and equivalents.

[0023] A speaker change detection (SCD) system consistent with the present invention performs speaker change detection on an input audio stream. The speaker change detection can be performed at real-time or faster than real-time. Speaker changes are detected only at phone boundaries detected within audio segments of a predetermined length. The predetermined length of the audio segments may be set at a much longer length, such as 30 seconds, than conventional predetermined intervals (e.g., 10 ms) used in detecting speaker changes.

EXEMPLARY SYSTEM

[0024] Fig. 2 is a diagram of an exemplary system 200 in which systems and methods consistent with the present invention may be implemented. In general, system 200 provides indexing and retrieval of input audio for clients 250. For example, system 200 may index input speech data, create a structural summarization of the data, and provide tools for searching and browsing the stored data.

[0025] System 200 may include multimedia sources 210, an indexer 220, memory system 230, and server 240 connected to clients 250 via network 260. Network 260 may include any type of network, such as a local area network (LAN), a wide area network (WAN) (e.g., the Internet), a public telephone network (e.g., the Public Switched Telephone Network (PSTN)), a virtual private network (VPN), or a combination of networks. The various connections shown in Fig. 2 may be made via wired, wireless, and/or optical connections.

[0026] Multimedia sources 210 may include one or more audio sources, such as radio broadcasts, or video sources, such as television broadcasts. Indexer 220 may receive audio data from one of these sources as an audio stream or file.

[0027] Indexer 220 may receive the input audio data from multimedia sources 210 and generate a rich transcription therefrom. For example, indexer 220 may segment the input data by speaker, cluster audio segments from the same speaker, identify speakers by name or gender, and transcribe the spoken words. Indexer 220 may also segment the input data based on topic and locate the names of people, places, and organizations. Indexer 220 may further analyze the input data to identify when each word was spoken (possibly based on a time value). Indexer 220 may include any or all of this information as metadata associated with the transcription of the input audio data. To this end, indexer 220 may include speaker segmentation logic 221, speech recognition logic 222, speaker clustering logic 223, speaker identification logic 224, name spotting logic 225, topic classification logic 226, and story segmentation logic 227.

[0028] Speaker segmentation logic 221 detects changes in speakers in a manner consistent with aspects of the present invention. Speaker segmentation logic 221 is described in more detail below.

[0029] Speech recognition logic 222 may use statistical models, such as acoustic models and language models, to process input audio data. The language models may include n-gram language models, where the probability of each word is a function of the previous word (for a bi-gram language model) and the previous two words (for a tri-gram language model). Typically, the higher the order of the language model, the higher the recognition accuracy at the cost of slower recognition speeds. The language models may be trained on data that is manually and accurately transcribed by a human.

[0030] Speaker clustering logic 223 may identify all of the segments from the same speaker in a single document (i.e., a body of media that is contiguous in time (from beginning to end or from time A to time B)) and group them into speaker clusters. Speaker clustering logic 223 may then assign each of the speaker clusters a unique label. Speaker identification logic 224 may identify the speaker in each speaker cluster by name or gender.

[0031] Name spotting logic 225 may locate the names of people, places, and organizations in the transcription. Name spotting logic 225 may extract the names and store them in a database. Topic classification logic 226 may assign topics to the transcription. Each of the words in the transcription may contribute differently to each of the topics assigned to the transcription. Topic classification logic 226 may generate a rank-ordered list of all possible topics and corresponding scores for the transcription.

[0032] Story segmentation logic 227 may change the continuous stream of words in the transcription into document-like units with coherent sets of topic labels and other document features. This information may constitute metadata corresponding to the input audio data. Story segmentation logic 227 may output the metadata in the form of documents to memory system 230, where a document corresponds to a body of media that is contiguous in time (from beginning to end or from time A to time B).

[0033] In one implementation, logic 222-227 may be implemented in a manner similar to that described by John Makhoul et al., "Speech and Language Technologies for Audio Indexing and Retrieval," Proceedings of the IEEE, Vol. 88, No. 8, August 2000, pp. 1338-1353, which is incorporated herein by reference.

[0034] Memory system 230 may store documents from indexer 220. Memory system 230 may include one or more databases 231. Database 231 may include a conventional database, such as a relational database, that stores documents from indexer 220. Database 231 may also store documents received from clients 250 via server 240. Server 240 may include logic that interacts with memory system 230 to store documents in database 231, query or search database 231, and retrieve documents from database 231.

[0035] Server 240 may include a computer or another device that is capable of interacting with memory system 230 and clients 250 via network 260. Server 240 may receive queries from clients 250 and use the queries to retrieve relevant documents from memory system 230. Clients 250 may include personal computers, laptops, personal digital assistants, or other types of devices that are capable of interacting with server 240 to retrieve documents from memory system 230. Clients 250 may present information to users via a graphical user interface, such as a web browser window.

[0036] Typically, in the operation of system 200, audio streams are transcribed as rich transcriptions that include metadata that defines information, such as speaker identification and story segments, related to the audio streams. Indexer 220 generates the rich transcriptions. Clients 250, via server 240, may then search and browse the rich transcriptions.

[0037] Fig. 3 is a diagram illustrating an exemplary computing device 300 that may correspond to server 240 or clients 250. Computing device 300 may include bus 310, processor 320, main memory 330, read only memory (ROM) 340, storage device 350, input device 360, output device 370, and communication interface 380. Bus 310 permits communication among the components of computing device 300.

[0038] Processor 320 may include any type of conventional processor or microprocessor that interprets and executes instructions. Main memory 330 may include a random access memory (RAM) or another type of dynamic storage device that stores information and instructions for execution by processor 320. ROM 340 may include a conventional ROM device or another type of static storage device that stores static information and instructions for use by processor 320. Storage device 350 may include a magnetic and/or optical recording medium and its corresponding drive.

[0039] Input device 360 may include one or more conventional mechanisms that permit an operator to input information to computing device 300, such as a keyboard, a mouse, a pen, a number pad, a microphone and/or biometric mechanisms, etc. Output device 370 may include one or more conventional mechanisms that output information to the operator, including a display, a printer, speakers, etc. Communication interface 380 may include any transceiver-like mechanism that enables computing device 300 to communicate with other devices and/or systems. For example, communication interface 380 may include mechanisms for communicating with another device or system via a network, such as network 260.

EXEMPLARY PROCESSING

[0040] As previously mentioned, speaker segmentation logic 221 locates positions in the input audio stream at which there are speaker changes. Fig. 4 is diagram illustrating functional components of speaker change detection logic 221. As shown, speaker change detection logic 221 includes input stream segmentation component 401, phone classification decode component 402, and speaker change detection (SCD) component 403.

[0041] Input stream segmentation component 401 breaks the input stream into a continuous stream of audio intervals. The audio intervals may be of a predetermined length, such as 30 seconds. Fig. 5 illustrates an audio stream broken into 30 second intervals 501-504. Two successive intervals may include overlapping portions. Intervals 502 and 503, for example, may include overlapping portion 505. Overlapping portion 505 ensures that speakers changes at the boundaries of two intervals can still be detected. Portion 505 is the end of audio interval 502 and the beginning of audio interval 503. Input stream segmentation component 401 passes successive intervals, such as intervals 501-504, to phone classification decode component 402.

[0042] Returning to Fig. 4, phone classification decode component 402 locates phones in each of its input audio intervals, such as intervals 501-504. A phone is the smallest acoustic event that distinguishes one word from another. The number of phones used to represent a particular language may vary depending on the particular phone model that is employed. In the phone system described in Kubala et al., "The 1997 Byblos System Applied to Broadcast News Transcription," *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop,* Landsdowne, VA, February 1998, 45 context independent phone models were trained for both genders. Another silence phone model was trained for silence, giving a total of 91 phone models. The 91 phone models were then used to decode speech, resulting in an output sequence of phones with gender and silence labels.

[0043] Phone classification decode component 402, instead of attempting to decode audio intervals 501-504 using a complete phone set, classifies audio intervals 501-504 using a simplified phone decode set. More particularly, phone classification decode component 402 may classify phones into three phone classes: vowels and nasals, fricatives or sibilants, and obstruents. The phones within each class have similar acoustic characteristics. Vowels and nasals are similar in that they both have pitch and high energy.

[0044] Phone classification decode component 402 may use four additional phones directed to non-speech events: music, laughter, breath and lip-smack, and silence. This gives a total of eight phone classes, which are illustrated in Table I, below.

Table I

CLASS	CONVENTIONAL PHONES INCLUDED IN CLASS
Vowels and Nasals	AX, IX, AH, EH, IH, OH,UH, EY, IY,AY, OY, AW. OW, UW, AO, AA, AE, EI, ER, AXR, M, N, NX, L, R, W, Y
Fricatives	V, F, HH, TH, DH, Z, ZH, S, SH
Obstruents	B, D, G, P, T, K, DX, JH, CH
Music	
Laughter	
Breath and lip-smack	
Silence	

[0045] As shown in Table I, the phone class “vowels and nasals” includes a number of conventional phones. Similarly, the phone classes “fricatives” and “obstruents” may also include a number of conventional phones. Phone classification decode component 402 does not, however, need to distinguish between the individual phones shown in any particular phone class. Instead, phone classification decode component 402 simply classifies incoming audio as a sequence of the phone classes shown in Table I. This allows phone classification decode component 402 to be trained on a reduced number of events (seven), thus requiring a simpler, and a more computationally efficient, phone decode model.

[0046] Phone classification decode component 402 may use a 5-state Hidden Markov Model (HMM) to model each of the seven phone classes. One codebook with 64 diagonal Gaussian Mixture Models (GMM) is shared by the 5 states from the same phone class and for each state a Gaussian mixture weight is trained. One suitable implementation of phone class decode component 402 is discussed in Daben Liu et al., "Fast Speaker Change Detection for Broadcast News Transcription and Indexing," *Proceedings of Eurospeech 99*, Budapest, Hungary, September 1999, pp. 1031-1034, which is incorporated herein by reference.

[0047] Fig. 6 is a diagram illustrating exemplary phone classes 601-605 decoded for an audio interval, such as interval 501, by phone classification decode component 402. A typical 30 second speech interval may include approximately 300 phone classes. Between each phone class is a phone boundary 610-612.

[0048] Returning to Fig. 4, SCD component 403 receives the phone-decoded audio intervals from phone classification decode component 402. SCD component 403 may then analyze each boundary 610-612 to determine if the boundary corresponds to a speaker change. More specifically, SCD component 403 compares the audio signal before the boundary and after the boundary within the audio interval. For example, when analyzing boundary 611 in interval 501, SCD component 403 may compare the audio signal corresponding to phone classes 601 and 602 with the audio signal corresponding to classes 603-605.

[0049] Fig. 7 is a flow chart illustrating the operation of SCD component 403 in additional detail when detecting speaker change boundaries in intervals 501-504 (Fig. 5).

[0050] To begin, SCD component 403 may set the starting position at one of the phone class boundaries of an interval to be analyzed (Act 701). The selected boundary may not be the first boundary in the interval as there may not be sufficient audio data between the start of the interval and the first boundary for a valid comparison of acoustic features. Accordingly, the first selected boundary may be, for example, half-way into overlapping portion 505. SCD component 403 may similarly process the previous interval up to the point half-way into the same overlapping portion. In this manner, the complete audio stream is processed.

[0051] SCD component 403 may next calculate cepstral vectors for the regions before and after the selected boundary of the active interval (Act 702). The calculation of cepstral vectors for samples of audio data is well known in the art and will not be described in detail herein. The two cepstral vectors are compared (Act 703). In one implementation consistent with the present invention, the two cepstral vectors are compared using the generalized likelihood ratio test. Assume that the cepstral vector to the left of the selected boundary is vector **x** and the cepstral vector to the right of the selected boundary is vector **y**. The generalized likelihood ratio test may be written as:

$$\lambda = \frac{L\left(z; u_z, \sum_z\right)}{L\left(x; u_x, \sum_x\right) L\left(y; u_y, \sum_y\right)},$$

where $L\left(v; u_v, \sum_v\right)$ is the maximum likelihood of **v** and **z** is the union of **x** and **y**.

[0052] If λ is above a predetermined threshold, the two vectors are considered to be similar to one another, and are assumed to be from the same speaker (Acts 704 and 705). Otherwise, the two vectors are dissimilar to one another, and the boundary point corresponding to the two vectors is defined as a speaker change boundary (Acts 704 and 706). The predetermined threshold may be determined empirically.

[0053] If there are any further boundaries in the interval, SCD component 403 repeats Acts 702-706 for the next boundary (Acts 707 and 708). If there are no further boundaries, the interval has been completely processed.

[0054] In alternate implementations, SCD component 403 may apply different threshold levels for boundaries that surround the non-speech phones. Thus, in between speech phones, SCD component 403 may use a threshold that is less likely to give a speaker boundary change indication than a boundary between non-speech phones. Further, it is often the case that λ tends to be larger when calculating for larger data sets. Accordingly, a bias factor based on the size of the data set may be added to λ to compensate for this property.

CONCLUSION

[0055] Speaker segmentation logic, as described herein, detects changes in speakers at boundary points determined by the location of phones in speech. The phones are decoded for the speech using a reduced set of possible phones. Further, the speaker segmentation logic processes the audio as discrete intervals of audio in which boundary portions of the audio are set to overlap to ensure accurate boundary detection over the whole audio signal.

[0056] The foregoing description of preferred embodiments of the invention provides illustration and description, but is not intended to be exhaustive or to limit the invention to the precise form disclosed. Modifications and variations are possible in light of the above teachings or may be acquired from practice of the invention. Moreover, while a series of acts have been presented with respect to Fig. 7, the order of the acts may be different in other implementations consistent with the present invention.

[0057] Certain portions of the invention have been described as software that performs one or more functions. The software may more generally be implemented as any type of logic. This logic may include hardware, such as application specific integrated circuit or a field programmable gate array, software, or a combination of hardware and software.

[0058] No element, act, or instruction used in the description of the present application should be construed as critical or essential to the invention unless explicitly described as such. Also, as used herein, the article “a” is intended to include one or more items. Where only one item is intended, the term “one” or similar language is used.

[0059] The scope of the invention is defined by the claims and their equivalents.